

Assessing the use of data from non-NOAA sources

A document prepared by the
Data Archiving and Access Requirements Working Group
(DAARWG)

Report to the
Science Advisory Board
National Oceanic and Atmospheric Administration

Presented by
Ferris Webster, chair, DAARWG

29 November 2011

A policy on using external data

- **NOAA's activities in observing, analysis, prediction, and response** will involve the use of external data from governmental and non-governmental data sources.
- The long-term objective of creating a national climate portal will involve the cooperative use of large data sets.
- The DAARWG document outlines issues to consider in developing a NOAA policy on using external data.

The benefits of a policy

- A timely NOAA policy for using external data **could improve NOAA's data activities and serve** as a model for adoption by partners.
 - The DAARWG believes that NOAA has a leadership role to play.
- We are in a period of rapid innovation in information-handling techniques and public acceptance of them.
- Having a policy may assist NOAA to flexibly assess the value of using external data as conditions change.

Policy Guidelines

- DAARWG presents a series of policy elements:
 1. Need for external data
 2. Life-cycle costs
 3. Data rights
 4. Data retention
 5. Data source
 6. Documentation
 7. Discovery and access
 8. System requirements
- For each element the document presents:
 - A brief statement of the DAARWG view.
 - Questions to be incorporated into a NOAA policy.

1 Need for external data

- Need should be paramount.
- NOAA should establish that there is truly a need for the external data.
- In emergency cases, especially when life or property is at risk, when NOAA has identified an external source of data that could significantly improve its forecasts, warnings, or analysis, NOAA should make every effort to obtain such data and to make it available to the public.

2 Life-cycle costs

- NOAA should consider life-cycle costs and evaluate cost-benefit before obtaining data from external sources.
- Life-cycle costs are frequently overlooked or underestimated.
- **Even “free” data encumber long-term life-cycle costs and obligations. There are no “free” data.**
- If benefits do not exceed costs, alternative procedures for acquiring the needed data should be explored.

3 Data rights

- NOAA should make its data open and available to any government, public, or private entity, provided without restriction on use and without limitation for further distribution.
- However, external datasets often come with restrictions on use by NOAA or on redistribution to others.
- NOAA should strive to make the data it obtains from external sources available to the public without restriction except when the proprietary rights of the data provider outweigh the public interest in having unrestricted access.

4 Data retention

- NOAA normally archives its datasets.
- NOAA should clarify long-term archiving needs before accepting external datasets.
- However, with the possible development of one or more federated data systems, large datasets may be too expensive to move, and may even be so big that it is impractical to host all data in one location.
- The result may be distributed-data architecture, with a federated storage system that will require cooperative archiving arrangements.

5 Source

- NOAA should ensure that data come from a certified and reliable source.
- A procedure for certifying data sources should be established.
- Uncertainty with regard to the data or their source should be documented.
- If NOAA uses external data for producing products and services that later turn out to be **unreliable, NOAA's credibility and reliability may be damaged.**

6 Documentation

- Data documentation (metadata) must be adequate for immediate and future use.
- Data provenance and the procedures used to develop products and services from data, should be transparently evident.
- Data may get used for a range of new things and current metadata may not be adequate.
- Recalibration of much early satellite data from the 1970s and 1980s reveals that information recorded with them is wholly inadequate.

7 Discovery & access

- For external datasets that are subject to ongoing retrieval, the system for finding and obtaining the data should be standardized to assure reliable access.
- For ongoing retrieval of data from an external source, data should be accessible via a standardized protocol in a well-known format.
- For operational access, the data source should be operationally reliable, with high availability, and redundancy.

8 System requirements

- Software and hardware should meet the needs for obtaining and ingesting external data.
- If accessing the data requires a NOAA receiving system, the system should have the necessary bandwidth and storage capacity.
- NOAA Personnel should be adequate to meet the incremental requirements of handling external datasets.

Conclusions

- DAARWG endorses the objective to create and implement a NOAA-wide policy for the use of external environmental data.
- The development of federated data systems will allow a more holistic and global approach to environmental data and services.
- In preparation for that development, NOAA will benefit by preparing a policy on using external data.
- With the experience gained with that policy, NOAA will be positioned to play its leadership role as national and international collaborative data systems develop.

DAARWG membership slides

- With rotation, one third of the DAARWG needs to be replaced each year.
- Suggestions for new members are always welcome.
- Look for expertise—diversity in membership is a strength.

DAARWG terms

Name (current term end date)	2011	2012	2013
Phil Jones	12/2011		
Roger Wakimoto	12/2011		
Bruce Wielicki	12/2011		
Dean Williams	12/2011		
Roberta Balstad		12/2012	
Peter Cornillon		12/2012	
Sara Graves		12/2012	
Ernest Hildner		12/2012	
Ferris Webster		12/2012	
John Boreman			12/2013
Kris Narasimham			12/2013
Daphne Fautin			12/2013
Chris Lenhardt			12/2013
Beth Plale			12/2013

Existing DAARWG members



- Roberta Balstad, 2012, Columbia University,
 - Societal impacts of climate change



- John Boreman, 2013, North Carolina State Univ.
 - Population dynamics and environmental impact assessments of coastal & estuarine fishes.



- Peter Cornillon, 2012, University of Rhode Island
 - Satellite oceanography, distributed data systems



- Sara Graves, 2012, U. of Alabama in Huntsville
 - Large-scale distributed information systems

Existing DAARWG members



- Ernest Hildner, 2012, Boulder, Colorado
 - Space environment research



- Phil Jones, 2011, University of East Anglia, UK
 - Climate change, paleoclimatology



- Krish Narasimham, 2013, Lockheed Martin Corporation
 - Enterprise IT, data management, communication networks.



- Roger Wakimoto, 2011, NCAR
 - Atmospheric sciences; severe weather

Existing DAARWG members



- Ferris Webster, 2012, U of Delaware
 - Physical oceanography, ocean data systems



- Bruce A. Wielicki, 2011, NASA Langley Research Center
 - Cloud layer properties



- Dean Williams, 2011, Lawrence Livermore National Laboratory
 - Climate modeling. IGBP model archive



- Thomas Zacharia, Oak Ridge National Laboratory
 - SAB Liaison

Existing DAARWG members



- Daphne Fautin, 2013, University of Kansas
 - Biodiversity, invertebrate biology, marine organisms



- Chris Lenhardt, 2013, Oak Ridge National Laboratory
 - Earth and environmental sciences, ecosystem science, environmental data science & systems



- Beth Plale, 2013, Indiana University
 - data-driven computing, long-term preservation of digital data, collaboration with earth and atmospheric sciences.

WG Terms of Reference

- Provide scientific advice and broad direction to NOAA regarding the wide range of data, information, and products that NOAA should archive and how NOAA can best provide access to this information.
- The Data Archiving and Access Requirements (DAAR) Working Group will evaluate data archiving and access requirements from all of NOAA's observing systems and computational models, as well as non-NOAA information.